# Countability and Number in Japanese-to-English Machine Translation

Francis Bond, Kentaro Ogura, Satoru Ikehara
**NTT Communication Science Laboratories**
1-2356 Take, Yokosuka-shi, Kanagawa-ken, Japan 238-03
{bond,ogura,ikehara}@nttkb.ntt.jp

COLING 94, August 1994*

## Abstract

This paper presents a heuristic method that uses information in the Japanese text along with knowledge of English countability and number stored in transfer dictionaries to determine the countability and number of English noun phrases. Incorporating this method into the machine translation system **ALT-J/E**, helped to raise the percentage of noun phrases generated with correct use of articles and number from 65% to 73%.

## 1 Introduction

Correctly determining number is a difficult problem when translating from Japanese to English. This is because in Japanese, noun phrases are not normally marked with respect to number. Japanese nouns have no equivalent to the English singular and plural forms and verbs do not inflect to agree with the number of the subject (Kuno 1973). In addition, there is no grammatical marking of countability.[1]

In order to generate English correctly, it is necessary to know whether a given noun phrase is countable or uncountable and, if countable, whether it is singular or plural. Deciding this is a problem even for humans translating from Japanese to English, but they have their own knowledge of both languages to draw on. A machine translation system needs to have this knowledge codified in some way. As generating articles and number is only important when the rest of the sentence has been correctly generated, there has not been a lot of research devoted to it. Recently, Murata and Nagao (1993) have proposed a method of determining the referentiality property and number of nouns in Japanese sentences for machine translation into English, but the research has not yet been extended to include the actual English generation.

This paper describes a method that extracts information relevant to countability and number from the Japanese text and combines it with knowledge about countability and number in English. First countability in English is discussed at the noun phrase and then the noun level. As a noun phrase's countability in English is affected by its referential property (generic, referential or ascriptive) we present a method of determining the referential use of Japanese noun phrases. Next the process of actually determining noun phrase countability and number is described. This is followed by some examples of sentences translated by the proposed method and a discussion of the results.

The processing described in this paper has been implemented in NTT Communication Science Laboratories' experimental machine translation system **ALT-J/E** (Ikehara *et al.* 1991). Along with new processing for the generation of articles, which is not discussed in detail in this paper, it improved the percentage of noun phrases with correctly generated determiners and number from 65% to 73%.

*This paper was presented at COLING 94 and appears in the proceedings: Vol I, pp 32–38.

[1] Japanese does not have obligatory plural morphemes. Plurality can be marked but only rarely is, for example by adding a suffix such as *tachi* "and others" (this can normally only be used with people or animals).

Table 1: Lexical information for nouns

| Noun | | Countability | Default | Default |
|---|---|---|---|---|
| English | Japanese | Preference | Number | Classifier |
| knife | *houchou* | FULLY COUNTABLE | SINGULAR | — |
| noodles | *men* | FULLY COUNTABLE | PLURAL | — |
| group | *mure* | (COLLECTIVE) | SINGULAR | — |
| cake | *ke–ki* | STRONGLY COUNTABLE | SINGULAR | — |
| beer | *bi–ru* | WEAKLY COUNTABLE | SINGULAR | — |
| furniture | *kagu* | UNCOUNTABLE | SINGULAR | *piece* |
| knowledge | *chishiki* | (SEMI-COUNTABLE) | SINGULAR | *piece* |
| scissors | *hasami* | PLURALIA TANTUM | PLURAL | *pair* |
| clothes | *ifuku* | PLURALIA TANTUM | PLURAL | — |

# 2 Countability

## 2.1 Noun Phrase Countability

We adopt the definition of countability in English given in Allan (1980:541–3). A countable noun phrase is defined as follows:

**I** If the head constituent of an NP falls within the scope of a denumerator it is countable.

**II** If the head constituent of an NP is plural it is countable.

Where "the phrase 'falls within the scope [or domain] of a denumerator' means 'is denumerated' by it; i.e the NP reference is quantified by the denumerator as a number of discrete entities."

Not all nouns in English can become the head of a countable noun phrase. In particular, noun phrases whose heads fall within the scope of a denumerator ('denumerated' noun phrases) must be headed by a noun that has both singular and plural forms. Nouns that do not have both forms, like *equipment* or *scissors*, require a classifier to be used. The classifier becomes the head of a countable noun phrase with the original noun attached as the complement of a prepositional phrase headed by *of*: *a pair of scissors, a piece of equipment.*

Whether a noun can be used to head a countable noun phrase or not depends both on how it is interpreted, and on its inherent countability preference. Noun countability preferences are discussed in the next section.

## 2.2 Noun Countability Preferences

A noun's countability preference determines how it will behave in different environments. We classify nouns into seven countability preferences, five major and two minor, as described below.

The two most basic types are 'fully countable' and 'uncountable'. Fully countable nouns, such as *knife* have both singular and plural forms, and cannot be used with determiners such as *much.*[2] Uncountable nouns, such as *furniture*, have no plural form, and can be used with *much.*

Between these two extremes there are a vast number of nouns, such as *cake*, that can be used in both countable and uncountable noun phrases. They have both singular and plural forms, and can also be used with *much.* Whether such nouns will be used countably or uncountably depends on whether their referent is being thought of as made up of discrete units or not. As it is not always possible to explicitly determine this when translating from Japanese to English, we divide these nouns into two groups: 'strongly countable', those that are more often used to refer to discrete entities, such as *cake*, and 'weakly countable', those that are more often used to refer to unbounded referents, such as *beer.*

The last major type of countability preference is 'pluralia tanta': nouns that only have a plural form, such as *scissors.* They can neither be denumerated nor modified by *much.* We fur-

[2] The determiners *much, little, a little, less and overmuch.* can all be used for this test

ther subdivide pluralia tanta into two types, those that can use the classifier *pair* to be denumerated, such as *a pair of scissors* and those that can't, such as *clothes*. 'pair' pluralia tanta have a singular form when used as modifiers (*a scissor movement*). Pluralia tanta such as *clothes*, use the plural form even as modifiers (*a clothes horse*), and need a countable word of similar meaning to be substituted when they are denumerated: *a garment, a suit, ...*.

The two minor types are subsets of fully countable and uncountable nouns respectively. Unless explicitly indicated, they will be treated the same as their supersets. 'Collective' nouns share all the properties of fully countable nouns. In addition they can have singular or plural verb agreement with the singular form of the noun: *The government has/have decided.* 'Semi-countable' nouns share the properties of uncountable nouns, except that they can be modified directly by *a/an*; for example *a knowledge [of Japanese]*.

Examples of the information about countability and number stored in the Japanese to English noun transfer dictionary are given in table 1. The information about noun countability preferences cannot be found in standard dictionaries and must be entered by an English native speaker. Some tests to help determine a given noun's countability preferences are described in Bond and Ogura (1993), which discusses the use of noun countability preferences in Japanese to English machine translation.

## 3 Determination of NP Referentiality

The first stage in generating the countability and number of a translated English noun phrase is to determine its referentiality. We distinguish three kinds of referentiality: 'generic', 'referential' and 'ascriptive'.

We call noun phrases used to make general statements about a class generic; for example *Mammoths are extinct*. The way generic noun phrases are expressed in English is described in Section 3.1. Referential noun phrases are ones that refer to some specific referent; for example *Two dogs chase a cat*. Their number and countability are ideally determined by the properties of the referent. Ascriptive noun phrases are used to ascribe a property to something; for example *Hathi is an elephant*. They normally have the same number and countability as the noun phrase whose property they are describing.

1. if restrictively modified then 'referential'
   *my book, the man who came to dinner*
2. if subject of *extinct, evolve* ...'generic'
   *Mammoths are extinct*
3. if the semantic category of the subject of a copula is a daughter of the semantic category of the object then 'generic'
   *Mammoths are animals*
4. if modified by *aimed at, for* ...then 'generic'
   *A magazine for women*
5. if object of *like* ...then 'generic'
   *I like cake*
6. if complement of a copula then 'ascriptive'
   *NTT is a telephone company*
7. if appositive then 'ascriptive'
   *NTT, a telephone company* ...
8. default 'referential'

Figure 1: Determination of NP referentiality

The process of determining the referentiality of a noun phrase is shown in Figure 1. The tests are processed in the order shown. As far as possible, simple criteria that can be implemented using the dictionary have been chosen. For example, Test 4 " if a NP is modified by *aimed at, for* ...then it is 'generic'" is applied as part of translating NP1-*muke* into "for NP1". The transfer dictionary includes the information that in this case, NP1 should be generic.

Tests 2 a3 show two more heuristic methods for determining whether a noun phrase has generic reference. In Test 2, if the predicate is marked in the dictionary as one that only applies to classes as a whole, such as *evolve* or *be extinct*, then the sentence is taken to be generic. In Test 3, **ALT-J/E**'s semantic hierarchy is used to test whether a sentence is generic or not. For example in *Mammoths are animals*, *mammoth* has the semantic category ANIMAL so the sentence is judged to be stating a fact true of all mammoths and is thus generic.

### 3.1 Generic noun phrases

A generic noun phrase (with a countable head noun) can generally be expressed in three ways (Huddleston 1984). We call these GEN 'a', where the noun phrase is indefinite: *A mammoth is a mammal*; GEN 'the', where the noun phrase is definite: *The mammoth is a mammal*; and GEN $\phi$, where there is no article: *Mammoths are mammals*. Uncountable nouns and pluralia tanta can only be expressed by GEN $\phi$ (eg: *Furniture is expensive*). They cannot take GEN 'a' because they cannot be modified by *a*. They do not take GEN 'the', because then the noun phrase would normally be interpreted as having definite reference. Nouns that can be either countable or uncountable also only take GEN $\phi$: *Cake is delicious/Cakes are delicious*. These combinations are shown in Table 2, noun phrases that can not be used to show generic reference are marked *.

Table 2: Genericness and Countability

| GEN | Noun Countability Preference | | |
|---|---|---|---|
| type | Countable | Both | Uncountable |
| 'a' | a mammoth | *a cake | *a furniture |
| 'the' | the mammoth | *the cake | *the furniture |
| $\phi$ | mammoths | cake/cakes | furniture |

The use all three kinds of generic noun phrases is not acceptable in some contexts, for example * *a mammoth evolved*. Sometimes a noun phrase can be ambiguous, for example *I like the elephant*, where the speaker could like a particular elephant, or all elephants.

Because the use of GEN $\phi$ is acceptable in all contexts, **ALT-J/E** generates all generic noun phrases as such, that is as bare noun phrases. The number of the noun phrase is then determined by the countability preference of the noun phrase heading it. Fully countable nouns and pluralia tanta will be plural, all others are singular.

## 4 Determination of NP Countability and Number

The following discussion deals only with referential and ascriptive noun phrases as generic noun phrases were discussed in Section 3.1,

1. if the Japanese is explicitly plural then countable and plural
2. determine according to determiner
   *one dog, all dogs*
3. determine according to classifier
   *a slice of cake, a pile of cakes*
4. determine according to complement
   *schools all over the country*
5. ascriptive NPs match their subjects
   *A computer is a piece of equipment*
6. determine according to verb
   *I gather flowers*
7. use default value
   (a) uncountable, weakly countable become:
   uncountable and singular
   (b) pluralia tanta become:
   countable and plural
   (c) countable and strongly countable become:
   countable and singular or plural
   according to the dictionary default

Figure 2: Determination of English noun phrase Countability and Number

Table 3: Noun Phrase Countability and Number

| Noun Type | Denumerated | | Mass | |
|---|---|---|---|---|
| | Singular | Plural | Countable | Uncountable |
| Fully Countable | a dog | two dogs | dogs | dogs |
| Strongly Countable | a cake | two cakes | cakes | cake |
| Weakly Countable | a beer | two beers | beer | beer |
| Uncountable | a piece of information | two pieces of information | information | information |
| Pluralia Tantum | a pair of scissors | two pairs of scissors | scissors | scissors |

The definitions of noun phrase countability given in Section 2, while useful for analyzing English, are not sufficient for translating from Japanese to English. This is because in many cases it is impossible to tell from the Japanese form or syntactic shape whether a translated noun phrase will fall within the scope of a denumerator or not. Japanese has no equivalent to *a/an* and does not distinguish between countable and uncountable quantifiers such as *many/much* and *little/few*. Therefore to determine countability and generate number we need to use a combination of information from the Japanese original sentence, and default information from the Japanese to English transfer dictionary. As much as possible, detailed information is entered in the transfer dictionaries to allow the translation process itself to be made simple.

The process of determining a noun phrase's countability and number is shown in Figure 2. The process is carried out during the transfer stage so information is available from both the Japanese original and the selected English translation.

To make the task of determining countability and number simpler, we define combinations of different countabilities for nouns with different countability preferences that we can use in the dictionaries. The effects of the four most common types on the five major noun countability preferences are shown in Table 3.

Noun phrases modified by Japanese/English pairs that are translated as denumerators we call denumerated. For example a noun modified by *onoono-no* "each" is denumerated - singular, while one modified by *ryouhou-no* "both" is denumerated - plural. Uncountable and pluralia tantum nouns in denumerated environments are translated as the prepositional complement of a classifier. A default classifier is stored stored in the dictionary for uncountable nouns and pluralia tanta. Ascriptive noun phrases whose subject is countable will also be denumerated.

The two 'mass'[3] environments shown in Table 3 are used to show the countability of nouns that can be either countable or uncountable. Weakly countable nouns will only be countable if used with a denumerator. Strongly countable nouns will be countable and plural in such mass - countable environments as the object of *collect* (vt): *I collect cakes*, and uncountable and singular in mass -uncountable environments such as *I ate too much cake*. In fact, both *I collect cake* and *I ate too many cakes* are possible. As Japanese does not distinguish between the two the system must make the best choice it can, in the same way a human translator would have to. The rules have been implemented to generate the translation that has the widest application, for example generating *I ate too much cake*, which is true whether the speaker only ate part or all of one cake or if they ate many cakes, rather than *I ate too many cakes* which is only true if the speaker ate many cakes.

Sometimes the choice of the English translation of a modifier will depend on the countability of the noun phrase. For example, *kazukazu-no* and *takusan-no* can all be translated as "many". *kazukazu-no* implies that it's modificant is made up of discrete entities, so the noun phrase it modifies should be translated as denumerated - plural. *takusan-no* does not carry this nuance so **ALT-J/E** will translate a noun phrase modified by it as mass - uncountable, and *takusan-no* as *many* if the head is countable and *much* otherwise.

Rules that translate the nouns with different

[3] We called these environments 'mass' because they both can be used to show a mass or unbounded interpretation.

noun countability preferences into other combinations of countable and uncountable are also possible. For example, sometimes even fully countable nouns can be used in uncountable noun phrases. If an elephant is referred to not as an individual elephant but as a source of meat, then it will be expressed in an uncountable noun phrase: *I ate a slice of elephant.* To generate this the following rule is used: "nouns quantified with the classifier *kire* "slice" will be generated as the prepositional complement of *slice*, they will be singular with no article unless they are pluralia tanta, when they will be plural with no article".

Note that countable indefinite singular noun phrases without a determiner will have *a/an* generated. Countable indefinite plural noun phrases and uncountable noun phrases may have *some* generated; a full discussion of this is outside the scope of this article.

## 5 Experimental Results

This processing described above has been implemented in **ALT-J/E**. It was tested, together with new processing to generate articles, on a specially constructed set of test sentences, and on a collection of newspaper articles. The results are summarized in Table 4.

Table 4: Correct Generation of Articles and Number

| | Test Sentences | | Newspaper Articles | |
|---|---|---|---|---|
| | NPs (240) | Sentences (120) | NPs (717) | Sentences (102) |
| New | 94% | 90% | 73% | 12% |
| Old | 70% | 46% | 65% | 5% |

In the newspaper articles tested, there were an average of 7.0 noun phrases in each sentence. For a sentence to be judged as correct all the noun phrases must be correct. The introduction of the proposed method improved the percentage of correct sentences from 5% to 12%.

Some examples of translations before and after the introduction of the new processing are given below. The translations before the proposed processing was implemented are marked OLD, the translations produced by **ALT-J/E** using the proposed processing are marked NEW.

(1) *taitei-no kodomo-ha otona-ni naru*
most child adult become
OLD: "Most children become an adult"
NEW: "Most children become adults"

In (1), the noun phrase headed by *otona* "adult" is judged to be prescriptive, as it is the complement of the copular *naru* "become". Therefore the proposed method translates it with the same number as the subject.

(2) *manmosu-ha zetumetu-shita*
mammoth died-out
OLD: "A mammoth died out"
NEW: "Mammoths died out"

*zetumetu* "die out", is entered in the lexicon as a verb whose subject must be generic. *manmosu* "mammoth" is fully countable so the generic noun phrase is translated as a bare plural.

(3) *tofu 3-chou, hasami 1-chou,*
tofu 3, scissors 1,
*houchou 2-chou-ga aru*
knife 2 is
OLD: "There are 3 piece tofu, 1 scissors, and 2 knives"
NEW: "There are 3 pieces of tofu, 1 pair of scissors and 2 knives"

The old version recognizes that a denumerated noun phrase headed by an uncountable noun *tofu* requires a classifier but does not generate the correct structure neither does it generate a classifier for the pluralia tanta *scissors*. The version using the proposed method does.

(4) *sore-ha dougu da*
that equipment is
OLD: "That is equipment"
NEW: "That is a piece of equipment"

As the subject of the copula *that* is countable it's complement is judged to be denumerated by the proposed method. As the complement is headed by an uncountable noun it must be embedded in the prepositional complement of a classifier.

There are three main problems still remaining. The first is that currently the rules for determining the noun phrase referentiality are insufficiently fine. We estimate that if referentiality could be determined 100% correctly then the percentage of noun phrases with correctly generated articles and number could be improved to 96% in the test set we studied. The remaining 4% require knowledge from outside the sentence being translated. The biggest problem is noun phrases requiring world knowledge that cannot be expressed as a dictionary default. These noun phrases cannot be generated correctly by the purely heuristic methods proposed here. The last problem is noun phrases whose countability and number can be deduced from information in other sentences. We would like to extend our method to use this information in the future.

## 6 Conclusion

The quality of the English in a Japanese to English Machine Translation system can be improved by the method proposed in this paper. This method uses the information available in the original Japanese sentence along with information about English countability at both the noun phrase and noun level that can be stored in Japanese to English transfer dictionaries. Incorporating this method into the machine translation system **ALT-J/E** helped to improve the percentage of noun phrases with correctly generated articles and number from 65% to 73%.

## References

Allan, Keith. 1980. Nouns and countability. *Language* 56.541–67.

Bond, Francis, and Kentaro Ogura. 1993. Determination of whether an English noun phrase is countable or not using 6 levels of lexical countability. In *Proceedings of the 46th Annual Convention IPSJ Japan*, 6:107–108. (in Japanese).

Huddleston, Rodney. 1984. *Introduction to the Grammar of English*. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Proceedings of MT Summit III*, 101–106. (cmp-lg/9510008).

Kuno, Susumu. 1973. *The Structure of the Japanese Language*. Cambridge, Massachusetts, and London, England: MIT Press.

Murata, Masaki, and Makoto Nagao. 1993. Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-93)*, 218–25.